日本フランス語フランス文学会

１８世紀研究会：「日本における『百科全書』パリ版デジタルアーカイブ共同研究の現状」

メタデータをめぐる現状と今後の課題

報告者：小関武史（一橋大学）
cj00531@srv.cc.hit-u.ac.jp

要旨

- 『百科全書』パリ版デジタルアーカイブ共同研究の目的は、『百科全書』を自分の研究で利用したいと希望する人が真正なテクストに容易に到達できるよう環境整備することである。
- 現在は、見出し語や執筆者などごく基本的な条件に基づく検索が第１巻についてのみ可能になっている。
- 現状に関してはまだまだ改善すべき点が多い。
- 今後、さらに参照文献など抽出の難しいデータを整備する必要がある。
- 現状の改善や今後の進展には大勢の協力が不可欠である。

１．共同研究の目的

『百科全書』という膨大なテクストから、読みたい項目（の一部）に行き着くには、それなりの修練が必要です。重要な項目ほど同名の項目が複数存在し、執筆者が誰かを見極めるには略号を知っていなければならず、世に「パリ版」と称して出回っているものには信用できないものが少なくない――こうした問題は『百科全書』を参考までに読むだけなら、あまり気に留めるほどのこともないでしょう。しかし、研究論文で『百科全書』を利用するなら、真正のパリ版のテクストに基づかなければなりません。『百科全書』に関する重要な研究はいくつかありますが、リチャード・シュワッブ（Richard Schwab）の *Inventory of Diderot's Encyclopédie* は『百科全書』の「取扱説明書」のような性格を有しており、その成果を知らずに『百科全書』を利用するのは危険ですらあります。とはいえ、Schwab の研究成果を理解するまで『百科全書』に触れてはならないというのも極端な議論です。非専門家（これは『百科全書』の専門家ではないというだけの意味で、他の分野では専門家であることを前提としていますが）が自然と Schwab の成果をふまえた形で『百科全書』の真正テクストを読めるような環境の出現が臨まれます。われわれが構築しようとしている『百科全書』パリ版デジタルアーカイブは、まさにそのような環境をウェブ上に出現させることを目指しています。

２．共同研究の現状

現在は、ごく基本的な検索が第１巻についてのみウェブ上で可能となっています。ごく基本的な検索というのは、以下の通りです。

(1) 見出し語（の一部）
(2) Schwab 番号
(3) 分類符号
(4) 開始場所
(5) 巻数
(6) 品詞と性
(7) 長さ
(8) 執筆者

このうち、見出し語と Schwab 番号については検索者が語句を記入して検索をかける方式となっています。残りはプルダウンメニューから選ぶ方式です。番号は 5000 ほどあり、それをプルダウンにするのは現実的でないから記入式になっているだけで、実質的には検索者が自由に条件を設定できるのは見出し語のみということになります。これ以後、それぞれの検索条件について検討し、ついで現在は提供されていないが将来含めるとよいと思われる検索条件の実現可能性を考えてみることにします。

(1) 見出し語（の一部）

最も分かりやすい検索方法です。見出し語のすべてが一致しなくてもよいので、途中までを打ち込めばいくつかの候補を出せます。前方一致なので、見出し語が A ou B という形式になっているとき、B に相当する語句をもとに検索できないという問題はあります。その他、見出し語の一部にイタリックやスモール・キャピタルが使用されている場合、検索結果を表示する段階ではそうした情報を取り込めないという問題もありますが、これについては原版デジタルイメージを閲覧できるので、重大な欠陥と考える必要はありません。これに限らず、原版デジタルイメージを見れば簡単に解決する問題については、とくに対処する必要はないと考えています。

見出し語を animal で検索すると三つの項目がヒットする。

(2) Schwab 番号

これは Schwab の *Inventory* を手許に置いている人にとって、一発で目的の項目に辿り着ける便利な機能です。ただし、一般的にはあまり使用されないかもしれません。

他の条件で検索した結果が複数ある場合、それらは Schwab 番号をもとに並べられます。ただし、われわれの方式では Schwab が無視した項目も抽出したため、Schwab 番号に整数でないものが発生しました。そのせいもあるかもしれませんが、Schwab 番号が数値でなく文字列として処理されているようです。その結果、４桁の 1811 番よりもあとに３桁の 897 番が来てしまいます。一番上の位が小さいものから順に並ぶわけです。検索結果が少なければとくに問題はないのですが、多くの検索結果が表示されるときには支障が出そうです。

分類符号 Orfèvrerie で検索すると、1811 の ALLIANCE、1882 の ALOI の後に 897 の ADOUCIR が来てしまう。

(3) 分類符号

分類符号は整理が最も厄介な検索条件で、今後も整理が必要です。分類符号はディドロたち編集者が「百科全書的秩序」を整えるために必要不可欠と考えていた要素で、これがなければ『百科全書』はただの事典になってしまいます。つまり副題の Dictionnaire raisonné としての機能しか果たしません。分類符号の効果で項目相互の連関が生まれることによって、はじめて語の本来の意味における百科全書となるのです。それほど重要な役割を担っているにもかかわらず、分類符号の付け方は実に恣意的あるいは無秩序で、そのときどきで思いついた名称を分類符号として使っているとしか思えないほどです。その結果、同一の学問・技術（science ou art）分野を示すのに、似て非なる名称が乱立する羽目になりました。たとえば、大工の仕事に関することは charpenterie と charpentier(s)という二種類の分類符号を割り振られています。もし、これをオリジナル表記のままにしていれば、検索者は実際上不便を感じるにちがいありません。charpentier で検索をかけて一通り調べ、その後でまた charpenterie で調べ直すのは面倒です。かといって、見出し語のときのように検索ウィンドウを自由記述にして、charpent までの前方一致で検索するというのも、この事例だけを取れば有効ですが、『百科全書』全体で考えれば効果のほどは疑問です。なぜなら、『百科全書』で使用されている分類符号名称には思いも寄らないものが含まれており、そうした稀な分類符号を思いつくこと自体が困難だからです。見落としを避けるためにも、プルダウンメニュー形式を維持した方がよいでしょう。

以上のような事情を勘案すると、管理者が分類符号の表記統一を図るのが最善であるというのが、われわれの結論です。とはいえ、表記の統一を図るのは思いの外困難な作業でした。一例を挙げると、学

問分野を表すのに、charpenterie のような分野そのものを表す名詞か、それとも charpentiers のようなその分野に携わる人を表す名詞か、いずれを採用するかを考えたとき、『百科全書』の精神からすると前者が望ましいのは、多くの人が同意することと思います。しかし、たとえば花火職人を指す artificier に対応する artificerie という単語は存在せず、このようなときは artificier で我慢するしかありません。分類としての整合性は保てませんが、オリジナルの無秩序をある程度整理することまでしかできないのです。もちろん、このような表記の訂正は、管理者による判断に基づいて行われるので、オリジナル表記がどうなっていたかを分かるようにするのが、学問的に誠実な態度でしょう。そこで、検索結果ウィンドウで「すべての情報を表示」というボタンを押すと、分類符号のオリジナル表記を閲覧できるようにしました。

項目 ABATAGE の分類符号は、オリジナル表記が terme de Charpentier で、それを Charpenterie に手直しした。

今後の課題は、表記の手直しの仕方が適切かどうかを見直すことと、簡単な綴りミスを直すことです。綴りミスについては、この欄だけの問題ではありません。もう一つ大きな検討課題は、複数の分類符号による検索を行いたいという要望をどの程度考慮に入れるか、です。現在は分類符号一つだけを選ぶようになっています。chimie と médecine の両方に関わる項目をピックアップしたい、という要望には今のところ応えられません。

(4) 開始場所

開始場所は、その名の通り項目が始まっているページをもとに検索するための欄です。これもプルダウンメニューになっており、第１巻の項目が始まっているページ番号のみが選択可能になっています。

第２巻以降のメタデータが組み込まれたとき、異なる巻の数字をどう処理するかという問題が生じます。また、「開始」という概念にとらわれず、任意の数字を入力して、そのページに存在する（そこから始まっていたり、まだ途中であったり、そこで終わったり）項目が表示されるようにしてほしいという要望もありえます。それに応えられるようにするには、システム側で面倒な計算式を用意しなければならないかもしれません。ひとたび原版デジタルイメージを呼び出せば、あとはフリッパーの機能で任意のページに飛べるので、それで十分と思ってもらえると助かります。

(5) 巻数

現在は第１巻のメタデータしか登録されていないので、あまり役には立っていませんが、将来的には補助的な検索条件として必須の欄となるでしょう。これもプルダウンメニューでない方がよいという希望が出るかもしれません。『百科全書』出版の経緯をふまえ、1765 年にまとめて出版された第８巻から第１７巻までの１０冊を対象として絞り込むことができるだけでも、価値は大きいと思います。

(6) 品詞と性

品詞と性を単独で検索条件とすることは考えられません。他の条件と組み合わせての検索となるでしょう。KDDE のページが完成してから気づいたのですが、これについても分類符号と同じような表記の統一が必要です。たとえば、男性名詞が s.m.や sub.m.に分かれていても無意味だからです。プルダウンメニューを見ると、一見同じものがいくつか並んでいるように見えます。たとえば「adject.」が三つありますが、二つ目のものは adject.のあとに半角スペースが一つ、三つ目のものは半角スペースが二つ入っているため、機械が別のものとして認識したのです。こういったミスは訂正しなければなりませんし、こういう入力ミスが出にくい作業方法を考えなければなりません。

品詞と性に関しては、プルダウンメニューの出し方そのものを考え直すべきかもしれません。男性名詞とだけ記されているものと、男性名詞としても女性名詞としても使用される単語とが、現在の仕組みでは別々に示されますが、そのせいで目指す項目に行き着けない可能性も捨て切れないからです。もっとも、これはどちらかというと小さな問題でしょう。

(7) 長さ

長さは Schwab の数値に従っています。ただし、明らかに Schwab に誤りがあるものは［ ］の中に入れました。そのため、これも数値としては扱いにくくなっています。

ありそうな検索方法としては、長さの範囲を指定して、その条件に見合う項目を検索するというものです。たとえば、「長さが 1.0 以上で神学に関する項目を探す」といった方法です。今はこうした要求に応えることはできません。今後の課題です。

(8) 執筆者

執筆者についても「ノーマル表記」が必要です。ほとんどの人は執筆者の略号など知らないからです。大文字のOがダランベールで大文字のSがルソーである、といった知識を前提とすることはできません。通常の名前で検索できるような作り込みをする必要があります。もちろん、検索者が確認できるように、「すべての情報を表示」したときには、オリジナル表記が分かるように設定しておくべきでしょう。

３．これから抽出すべきメタデータと今後の課題

去年の夏に鷲見洋一先生の呼びかけによって、『百科全書』第 1 巻のメタデータ抽出作業が始まりました。これまでに三十人ほどのボランティアの方から協力をいただいています。厚く御礼申し上げます。

さて、メタデータ抽出作業は「全員必須の作業」と「任意の作業」に分かれていました。必須分のメタデータが、これまで紹介した分に当たります。今後は任意分のメタデータの整備が必要です。それらのメタデータとは、以下の通りです。

(9)　本文項目への参照
(10) 図版への参照
(11) 本文に明示された典拠情報（標題）
(12) 本文に明示された典拠情報（著者）
(13) 引用文
(14) 図表
(15) 特殊活字

これらの具体的内容については、時間の制約もあって触れることができません。テクストを緻密に読み込むことでしか抽出できないものが多い（とくに典拠情報）とだけ申し上げておきます。そうしたメタデータの抽出には、大勢のみなさんの献身的な協力が欠かせません。よろしくお願い申し上げます。

現在提供のデジタルアーカイブは試用版という位置付けです。みなさんに実際に活用していただいたうえで、ご意見・ご感想を承り、今後の方針を決める際の参考にさせていただきます。